DOCUMENT RESUME

ED 189 138 TM 800 308

AUTHOR Holmes, Susan E.
TITLE Item Banking: An Alternative Approach to Title I Evaluation?
PUB DATE Apr 80
NOTE 32p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests: Elementary Education: Evaluation Methods: *Item Banks: *Measurement Techniques: Norm Referenced Tests: *Program Evaluation: *Scores
IDENTIFIERS *California Achievement Tests: Elementary Secondary Education Act Title I: Linear Models: Oregon: *Rasch Model: RMC Models

ABSTRACT

The purpose of the research was to explcre an alternative implementation strategy for Elementary Secondary Education Act Title I evaluatien Model A: that is, the use of Rasch calibrated item banks rather than traditional ncrm-referenced tests. A methcdclogy for obtaining normative informatien for such banks involved equating a normed test with selected subsets of bank items. Three conventional equating techniques and one Rasch model technique were used for the development cf cenversion tables. All four equating techniques yielded similar results, but the accuracy of the results was of concern. The largest discrepancies between atual and predicted perfcrmance were observed in the score range in which most Title I students would be expected to score. An application of the conversion tables developed through the Project to real Title I and non-Title I data was reported. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED189138

Item Banking: An Alternative Approach to Title I Evaluation?

Susan E. Holmes
CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. Holmes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A paper presented at the annual meeting of the American Educational Research Association, Boston, April 9, 1980.

TM 800308

ABSTRACT

The research reported here explores the use of Rasch calibrated item banks as an alternative implementation strategy for Title I evaluation Model A. A methodology for obtaining normative information for such banks involving equating a normed test with selected subsets of bank items is discussed. The results of three conventional equating techniques, two linear and one equipercentile, were found to be consistent with those obtained from a Rasch model approach, although the accuracy of all four sets of results for Title I students was a serious concern. An application of the conversion tables developed to actual Title I reading and mathematics data is reported.

Item Banking: An Alternative Approach to Title I Evaluation?[1]

Susan E. Holmes
CTB/McGraw-Hill

Since 1974, the United States Office of Education has attempted to implement Title I evaluation reporting procedures which are outcome oriented and provide for the aggregation of achievement data across all Title I projects. The system is intended to facilitate the collection and analysis of data in six areas: student participation, parent involvement, staffing, inservice training, cost and project impact. Major emphasis has been placed on project impact, particularly for those projects providing instruction in the basic skill areas of reading, language arts and mathematics.

Three models or designs have been proposed for the evaluation of project impact, defined here in terms of achievement gains resulting from Title I intervention (Tallmadge and Wood, 1978). All three models may be used in conjunction with either a criterion-referenced or norm-referenced achievement test, and each involves the pretesting and posttesting of Title I students. Observed posttest performance is compared to a "no treatment expectation" which represents an estimate of what that performance would have been without Title I intervention. Achievement gains are then determined and quantified in terms of a common metric called the Normal Curve Equivalent (NCE)

[1]The research reported here was conducted while the author was a staff member of the Oregon Department of Education and was jointly sponsored by that agency and the United States Office of Education (Contract No. 300-79-0019). A copy of the complete project report can be obtained through ERIC under the Holmes (1980) citation in the reference section - this paper.

scale. This normalized standard score scale has a mean of 50 and a standard deviation of 21.06. It was chosen because of its equal-internal nature and its exact correspondence to the familar percentile scale at the points one, 50 and 99 (Tallmadge, 1976).

Of particular interest here is Model A, referred to as the norm-referenced model. The basic assumption of this model is that the percentile rank of Title I students will remain constant in relation to the group on which the achievement test was normed from pretest to posttest if no Title I intervention occurred. Of the several proposed models, Model A is the one most widely adopted since it utilizes procedures already in place in many projects. It does not necessitate the identification of a local comparison group, making it less expensive to implement. In addition, it is computationally less complex than the other models.

Model A may be implemented in two ways. The first, referred to as Model A1, involves administering a normed achievement test to Title I students at pretest and posttest time. A no-treatment expectation is determined by assuming that Title I students will maintain their percentile rank with respect to the norm group from pretest to posttest without Title I intervention. Their observed posttest status is then compared to this no-treatment expectation, and an improvement in performance is assumed to reflect Title I impact.

A second implementation plan for Model A, referred to as Model A2, has also been proposed. This model involves the administration of a non-normed achievement test at pretest and posttest time. In addition, a normed test is administered at either of the two test times. By equating the normed and non-normed tests, normative data compiled for the

normed test can be used to derive a no-treatment expectation for the non-normed test. As in Model A1, observed posttest performance is then compared to the no-treatment expectation.

Model A2 would appear to offer some recourse to those districts which are unable to locate a normed achievement test which adequately reflects their Title I program objectives. However, this approach is not without its drawbacks. For example, should these objectives change significantly enough in the future to necessitate changes in the non-normed test, the entire equating process outlined in Model A2 would need to be repeated for the new non-normed test. There is an additional concern. If both the normed and non-normed tests actually measure the same thing, as they must if equating techniques are to be appropriately applied, it would seem more reasonable to administer only the normed test and use a Model A1 approach. The computational complexity of Model A2, coupled with the prospect of replicating the effort every few years, makes many local districts reluctant to consider this approach to Title I evaluation.

A third alternative for the implementation of Model A, proposed here as "Model A3," has been suggested. This approach attempts to provide local districts with more flexibility in selecting achievement tests for Title I evaluation while eliminating the need for replicating the test equating effort each time a new test is selected or an existing one modified. Model A3 involves the use of item banks developed through Rasch model item analysis techniques (Rasch, 1960; 1966; Rentz and Bashaw, 1975; Wright, 1977).

Over the past few years, interest in applying Rasch model technology in educational measurement has greatly increased. One advantage of this

approach to measurement lies in the fact that, once a Rasch calibrated item bank has been constructed, all possible subsets of items drawn from the bank are automatically equated (i.e., result in test information on a common metric). In order to utilize a Rasch calibrated item bank in Title I evaluation reporting, however, normative information must be available for these common metric scores. By determining the relationship between selected subsets of items drawn from the bank and a series of normed tests, the normative data compiled for the normed test can be associated with the scale underlying the item bank. Once this relationship is defined, any appropriately chosen subset of bank items could then be used to evaluate project impact, provided the items were administered at an appropriate time of the year. Theoretically, local districts could select one set of items to identify Title I eligible students, another to determine the no-treatment expectation (pretest) for this group, and yet another to evaluate growth (posttest). Since all possible subsets of items drawn from the bank are automatically equated, it has been argued that tests which more accurately reflect local curriculum can be constructed in this way, with the test equating process being performed only once.

The purpose of the research reported here was to explore Model A3 as an alternative Model A implementation strategy for Title I evaluation reporting. Specific Project objectives were to:

1. describe a methodology for equating a Rasch calibrated item bank and a normed achievement test,
2. evaluate the results obtained from a Rasch model approach to test equating with those of more conventional approaches,
3. produce tables which permit the conversion of item bank scores into NCE units, and

4. provide a prototype Title I evaluation report based on tests developed from a Rasch calibrated item bank.

METHODS

Sample

Over 6,000 students were involved in the Project, about 1,000 at each grade level. The sample was obtained by inviting two Oregon school districts to participate. Twenty-three schools accepted, and the sample consisted of all students enrolled in grades three through eight in these schools.

Instruments

In order to obtain normative information for the item bank scale, pairs of tests, one member having existing norms and the other being drawn from a Rasch calibrated item bank, were administered to the same sample and then equated. Two banks, one consisting of reading items and the other of mathematics items, were used (Northwest Evaluation Association, 1978). Each contained over 1,100 goal-referenced items which had been field tested on relatively large samples before inclusion in the bank. Four tests in each content area were selected from a series of overlapping tests previously developed from these banks (Forster and Doherty, 1978). The normed test selected was the California Achievement Tests (1977), Form C (CAT). The total reading and total mathematics subtests from Levels 13 through 18 served as the second member in each test equating pair.

Data Collection

All students were administered the entire CAT level appropriate for their grade placement since these data were collected as part of a routine district wide evaluation program. In addition, approximately half took an item bank reading test, while the remaining half took an item bank mathematics test. Data were collected by the classroom teachers during regularly scheduled class time within the CAT recommended norming period.

Item Bank to NCE Conversion Tables

Development of tables for the conversion of item bank achievement estimates into their NCE equivalents involved several steps. Initially, the CAT and item bank test pairs were equated by applying a Rasch model equating technique to the two sets of item response data (Holmes, 1980). This process defined an equated or "converted" CAT score for each item bank raw score. Next, an item bank achievement estimate was obtained for each item bank raw score by applying the UCON procedures outlined by Wright and Stone (1979, pages 142-143). The item difficulty estimates used here were those already recorded in the bank for these items. Then the percentile rank associated with each converted CAT score was determined through interpolation from the raw score to percentile rank tables provided by the CAT publishers (California Achievement Tests, 1978). Values for both the beginning and end of each grade were obtained. And last, each percentile rank was converted into an NCE unit, thus completing the link from bank achievement estimate to NCE unit.

RESULTS

Initial Data Screening

Although all item bank test items were evaluated for fit before initial inclusion in the banks, a re-analysis was performed using the Project data. In general, difficulty estimates based on the Project sample were found to be quite consistent with those recorded in the banks for the same items. In most cases only one or two estimates were found to differ significantly enough to warrent deletion of these items from further analyses. No similar evaluation of the CAT items was conducted since all these items were needed in order to use the CAT norming information.

A second kind of data screening, involving the examination of student response data, was also performed. All students responding to fewer than 35 percent of the items on either test in an equating pair were deleted from further analyses. This deletion criterion was arbitrarily selected in an attempt to exclude students for whom the test equating pairs were inappropriate. The number of students remaining at each grade is presented in Table 1, along with the test equating pair administered at that grade.

Insert Table 1
about here

Test Summary Statistics

Summary test statistics based on those students and items retained after initial data screening can be seen in Table 2. The number of items, total test mean, standard deviation and KR20 reliability estimate are

reported for each test. The raw score correlation between test equating pairs, along with the value of these correlations corrected for attenuation, appear in Table 3. It should be noted that the corrected values represent the maximum coefficient obtainable if measurement error were totally eliminated. These values may be further inflated by the underestimated reliability coefficients used in their computation.

Insert Tables 2 and 3
about here

Test Equating

Four equating procedures were applied to each set of raw score pairs. Two linear methods, one based on observed scores and one on estimated true scores, as well as two non-linear methods, equipercentile equating and a Rasch model technique, were utilized (Holmes, 1980). Figure 1 presents a graph of the raw score equating results yielded by the four methods for the grade three reading data. It should be noted that no smoothing techniques were applied to the equipercentile results.

Insert Figure 1
about here

In order to evaluate the accuracy of the four sets of equating results, an overall average absolute discrepancy was computed for each technique. This value was obtained by determining the absolute difference between each student's actual CAT raw score and the converted CAT score predicted by each of the four equating methods, and then averaging these values across

students for each method. Table 4 presents the overall average absolute discrepancy values observed.

Insert Table 4
about here

In addition, the distribution of the average absolute discrepancy values observed for each item bank raw score was generated in an attempt to examine the nature of the equating discrepancies associated with each method. In all cases, the larger average values occurred at the lower end of the raw score distributions, regardless of equating technique. While not reported here, these data can be found in Holmes (1980).

In order to determine whether the overall discrepancies reported in Table 4 differed significantly, a one-way repeated measures analysis of variance was performed at each grade using individual absolute discrepancy values as the dependent variable. The results of these analyses are summarized in Table 5. In addition, the proportion of variance accounted for by equating method is reported there. It should be noted that although some analysis indicated significant differences between the overall average absolute discrepancy values observed for each equating technique, the proportion of variance accounted for by equating method was never greater than .0013.

Insert Table 5
about here

Rasch Model Equating Bias

While the overall average absolute discrepancy values reported in Table 4 are useful in comparing the four different equating methods studied here, the actual distribution of these discrepancies is helpful in evaluating the equating bias introduced by any particular method. Figure 2 presents a graph of the discrepancies actually observed for the Rasch model equating technique when applied to the grade three reading data. Entries above the horizontal zero line represent underestimated scores (i.e., the score actually obtained by a student was higher than the converted CAT score predicted for him by the Rasch model equating results), while those values below the line represent overestimated scores. The general configuration displayed in Figure 2 is representative of those observed in both content areas and all grades.

Insert Figure 2
about here

The Pearson product-moment correlation between the discrepancies resulting from the Rasch model technique and their corresponding item bank raw scores was then computed. While the values were small (-.08 to -.39), all correlations were negative and significantly different from zero ($p < .05$).

Conversion Tables

One of the Project objectives was to develop tables which facilitate the conversion of item bank achievement estimates into the NCE units necessary for Title I evaluation reporting. One conversion table was developed for each grade in each content area for a total of 12 tables.

A prototype Title I evaluation report was also developed using the fall and spring achievement estimates previously reported by Portland School District 1J in its 1978-79 Title I report (Holmes and Van Cleave, 1979). The average fall and spring Title I NCE equivalents and gains determined from the CAT-based conversion tables developed by the Project are presented in Table 6. To assist in interpreting these NCE values, it should be noted that the NCE and percentile scales correspond exactly at the points one, 50, and 99. An NCE of 50, therefore, represents the national average for that grade, (i.e., being exactly at grade level). An NCE gain that is greater than zero, on the other hand, indicates that the amount of learning that occurred between the two testings was greater than would have been expected had there been no Title I intervention. In general, the larger the NCE gain, the more effective the Title I project, although no specific NCE criterion gain can be designated since cost effectiveness is a consideration. For the sake of comparison, the same information is also presented for non-Title I students.

Insert Table 6
about here

DISCUSSION

Methodological Considerations

It should be noted that the test equating design used, that of administering both tests to the same group of students, was a fairly strong one. Although counterbalancing the order of the two test administrations would have eliminated possible fatigue or practice effects, this approach was not feasible since the normed data were collected as part of district evaluation programs and was therefore beyond the author's control. The design adopted, however, represented a highly desirable testing procedure for an equating study.

It should be noted that the Project sample, while not identical to the CAT norming group, did demonstrate a consistent relationship to it. A comparison of the CAT raw score means and standard deviations reported in Table 2 for the Project sample with those reported for the CAT spring norming sample (California Achievement Tests, 1979) indicated that the performance of the Project sample was higher and more homogeneous than that of the CAT norming sample (Holmes, 1980). These differences were, however, consistent across grades and content areas. In addition, the average item bank achievement estimates computed for the Project sample were consistent with the spring values reported for similar Portland students (Holmes and Van Cleave, 1979), suggesting similar performance levels for these two groups.

One of the assumptions inherent in test equating is that both tests measure the same thing. Since only total test equating was considered in the Project, the size of the correlation between test equating pairs is of interest. The correlations reported in Table 3 were, in some cases, lower than might be desired. The larger corrected values suggested that a relationship between the pairs did exist, although these corrected correlations are probably somewhat inflated. An interesting observation related to the subtest correlations should be noted. These values, not reported here, were smaller than those appearing in Table 3; that is, item bank and CAT total test correlations were consistently higher than item bank and CAT subtest correlations. This result is not surprising since the size of a correlation is affected by the number of observations used in its computation. It does, however, lend some support to the total test equating approach employed.

Equating Comparisons

No absolute criterion for evaluating the accuracy of equating results exists. However, a comparison of the overall average absolute difference between scores predicted by a particular technique and those actually observed should facilitate some general comparisons among techniques. An analysis of the Project sample values reported in Table 4 indicated that, at least for this criterion, the four equating procedures studied appeared to work equally well. The effect size analysis summarized in Table 5 suggested that the differences among overall average discrepancies were due to large sample sizes rather than any real differences among equating techniques. In all cases, the amount of variance accounted for by equating technique was less than one percent.

The similarity among the equating results observed here has also been reported elsewhere in the literature. Rentz and Bashaw (1975) and Beard and Pettie (1979) reported consistent equating results for conventional and Rasch model techniques. One disturbing aspect of this finding, however, is the size of the overall absolute average discrepancy values obtained (Table 4). These values are reported in CAT raw score points, suggesting that on the average all four equating techniques predicted CAT scores that differed by from five to seven points from the score actually obtained. These discrepancies are alarmingly large since the CAT standard errors of measurement are roughly 3.5 points. In addition, the discrepancy values reported are smaller than might be expected since these values were determined for the same sample used to generate the equating results.

A further examination of the absolute discrepancy values by item bank raw scores indicated that the largest of these values occurred at the lower end of the raw score distribution. This was the case for all equating techniques and probably reflects the general instability of test scores in this range of the distribution due to the presence of guessing. (These data can be found in Holmes (1980).) This situation suggests using an equating model which includes a guessing parameter. Regardless of the reason, Title I students may be expected to score in this portion of the distribution.

Since a Rasch model equating technique was used to generate the Project conversion tables, the bias introduced by this particular technique was further examined. Actual discrepancy values, rather than their absolute values, were plotted by item bank raw score (Figure 2). These plots confirmed the results reported above; that is, large discrepancies occurred

at the lower end of the distribution. The significant negative correlations between these discrepancies and their corresponding item bank raw scores suggest that the Rasch model technique results in a systematic underestimation of low scores. They further suggest a systematic overestimation of high scores, although this relationship cannot be confirmed because of the negatively skewed raw score distributions.

Conversion Tables

Application of the CAT-based conversion tables developed by the Project to actual Title I data resulted in a pattern of decreasing fall and spring NCE equivalents across grades (Table 6). This pattern, evident in both content areas, suggested that while large NCE gains from fall to spring were occurring, these Title I students rapidly lost ground as they moved through the grades, beginning each year at successively lower levels.

Since large equating discrepancies were observed for scores in the lower end of the distribution, the CAT-based conversion tables were also applied to the data from non-Title students (Table 6). The fall and spring equivalents obtained seem reasonable, with small positive and negative NCE gains being observed. In light of these values and those reported above for Title I students to, it is tempting to attribute the large Title I gains to the positive effect of Title I intervention. However, the inaccuracy of results occurring in the lower end of the raw score distribution noted earlier generates serious doubts about this interpretation.

An additional concern, alluded to earlier, involves the logical inconsistency of a Model A3 approach to Title I evaluation. In order to implement this model, both the normed and non-normed tests (i.e., the item bank) must

measure an identical trait. Test equating becomes inappropriate if this is not the case, and the degree to which it is not will be reflected in the accuracy of the equating results. It would seem that the flexibility suggested by this approach, and also by that of Model A2, is questionable. If, in fact, both tests measure an identical trait, a Model A1 approach to Title I evaluation would seem most reasonable.

CONCLUSIONS

The purpose of the research reported here was to explore an alternative implementation strategy for Title I evaluation Model A, namely the use of Rasch calibrated item banks. A methodology for obtaining the norm-referenced information that would be necessary to utilize such a bank was described, and tables which facilitated these conversions were developed for grades three through eight in the content areas of reading and mathematics.

Three conventional equating techniques and one Rasch model technique were considered during the development of the conversion tables. While no absolute criterion for evaluating the accuracy of the four sets of results exists, the conventional techniques produced results which were similar to those of the Rasch model approach when comparisons were based on the predicted and observed performance of the Project sample.

Although all four equating techniques yielded similar results, the accuracy of those results, especially in the lower end of the score distribution, was a concern. The largest discrepancies between actual and predicted performance were observed in the score range in which most Title I students may be expected to score.

An application of the conversion tables developed through the Project to real Title I and non-Title I data was reported. Large fall to spring NCE gains were observed for the Title I data. Serious concern related to the accuracy of these results was expressed because of the large discrepancy values observed for scores in the lower end of the distribution. It was concluded that additional research was necessary before this approach to Title I evaluation reporting could be implemented.

REFERENCES

Beard, J.G. & Pettie, A.L. A comparison of linear and Rasch equating results for basic skills assessment tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

California Achievement Tests, Forms C & D. Monterey, California: CTB/McGraw-Hill, 1977.

California Achievement Tests: Norms Tables. Monterey, California: CTB/McGraw-Hill, 1978.

California Achievement Tests: Technical Bulletin 1. Monterey, California: CTB/McGraw-Hill, 1979.

Comprehensive Tests of Basic Skills. Monterey, California: CTB/McGraw-Hill, 1973.

Forster, F. & Doherty, V. Using the Rasch model to improve your district's testing program. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Holmes, J.N. & Van Cleave, M.E. Area I Disadvantaged Child Project, Evaluation report, ESEA, Title I, 1978-79. Portland, Oregon: Portland School District 1J, August 1979.

Holmes, S.E. ESEA Title I evaluation and reporting refinement, The Title I linking project, Final project report. Salem, Oregon: Department of Education, Research and Assessment Division, 1980.

Northwest Evaluation Association. Rasch model monograph series. Portland, Oregon: Author, December 1978.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institute, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Rentz, R.R. & Bashaw, W.L. Equating reading tests with the Rasch model, Volume I final report, Volume II technical reference tables. Athens, Georgia: University of Georgia, Educational Research Laboratory, 1975.

Tallmadge, G.K. Interpreting NCEs, ESEA Title I evaluation and reporting system. (Technical paper No. 2). Mountain View, California: RMC Corporation, 1976.

Tallmadge, G.K. & Wood, C.T. User's guide, ESEA Title I, Evaluation and reporting system. Mountain View, California: RMC Research Corporation, January 1978.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

REFERENCES

Wright, B.D. & Stone, M. H. Best test design, Rasch measurement. Chicago, Illinois: Mesa Press, 1979.

TABLE 1

Number of Students Tested by Content Area, Grade and Test Equating Pair

	EQUATING TEST PAIR			
	GRADE	CAT CODE	ITEM BANK CODE	NUMBER OF PAIRS[a]
READING	3	CAT13	318	463
	4	CAT14	319	457
	5	CAT15	320	481
	6	CAT16	320	462
	7	CAT17	321	479
	8	CAT18	321	559
MATHEMATICS	3	CAT13	818	343
	4	CAT14	819	354
	5	CAT15	820	370
	6	CAT16	820	429
	7	CAT17	821	549
	8	CAT18	821	559

[a]Sample sizes include only those students who responded to at least 35 percent of the items on both tests.

TABLE 2

Number of Test Items, Total Test Raw Score Mean, Standard Deviation and Reliability Estimate for Each Test Equating Pair Member

		ITEM BANK				CAT			
	GRADE	NUMBER of ITEMS[a]	MEAN	STANDARD DEVIATION	KR20	NUMBER of ITEMS	MEAN	STANDARD DEVIATION	KR20
READING	3	44	31.1	9.0	.92	73	59.2	11.0	.93
	4	42	27.4	7.6	.88	70	49.6	13.5	.94
	5	45	28.8	8.6	.89	70	51.7	12.1	.93
	6	42	28.7	8.2	.90	70	48.3	12.5	.93
	7	43	22.9	7.5	.84	70	46.8	12.3	.92
	8	44	26.2	8.1	.87	70	43.8	13.0	.93
MATHEMATICS	3	54	36.9	10.1	.91	85	56.5	14.6	.94
	4	53	34.8	11.0	.93	85	52.0	15.9	.94
	5	52	29.0	10.0	.90	85	51.9	15.5	.94
	6	48	33.1	8.9	.90	85	52.2	14.6	.93
	7	49	29.1	10.1	.91	85	48.6	16.5	.94
	8	47	31.5	9.6	.91	85	55.0	16.9	.95

[a]After initial item deletions

TABLE 3

Equating Test Pair Raw Score Correlations Uncorrected and Corrected for Attenuation

	READING		MATHEMATICS	
GRADE	$\underline{r}$	CORRECTED $\underline{r}$	$\underline{r}$	CORRECTED $\underline{r}$
3	.70	.76	.87	.94
4	.82	.90	.87	.93
5	.83	.91	.85	.92
6	.85	.93	.88	.96
7	.79	.90	.88	.94
8	.84	.93	.84	.90

TABLE 4

Overall Average Absolute Discrepancy Between Observed CAT Scores and the Converted CAT Scores Predicted by Each of the Four Equating Techniques (LO-Linear Observed; LT-Linear True; EQ-Equipercentile; RM-Rasch Model)

		EQUATING TECHNIQUE			
	GRADE	LO	LT	EQ	RM
READING					
	3	6.02	6.04	6.05	6.05
	4	6.24	6.37	5.97	5.89
	5	5.40	5.48	5.24	5.08
	6	5.26	5.31	5.27	5.29
	7	6.17	6.34	6.12	5.90
	8	5.68	5.77	5.67	5.53
MATHEMATICS					
	3	5.71	5.75	5.68	5.78
	4	6.45	6.47	6.30	6.48
	5	6.55	6.63	6.46	6.52
	6	5.81	5.85	5.77	6.10
	7	6.41	6.47	6.16	6.32
	8	6.89	6.96	6.96	6.84

TABLE 5

Summary of the One-way Repeated Measures Analysis of Variance Performed Using Absolute Discrepancy Between Observed and Converted CAT Scores as the Dependent Variable

	GRADE	F	df ERROR[1]	PROBABILITY	PROPORTION OF VARIANCE[2]
READING					
	3	.16	1365	.92	.0000
	4	25.42	1368	< .01	.0013
	5	16.11	1434	< .01	.0010
	6	1.30	1377	.27	.0000
	7	10.58	1434	< .01	.0010
	8	14.76	1674	< .01	.0003
MATHEMATICS					
	3	2.60	1020	.05	.0001
	4	2.10	1059	.10	.0002
	5	4.55	1107	< .01	.0001
	6	7.75	1281	< .01	.0009
	7	11.00	1638	< .01	.0005
	8	5.57	1659	< .01	.0001

[1]For all F-values reported the treatment degrees of freedom are 3.

[2]The ratio of treatment to total sum of squares.

TABLE 6

Average Title I and Non-Title I NCE Equivalents and Gains as Determined from CAT- Based Conversion Tables

		TITLE I				NON-TITLE I		
	GRADE[a]	N	FALL	SPRING	GAIN	N	FALL	SPRING	GAIN
READING	4	720	23.2	35.8	12.6	2660	56.9	55.0	-1.9
	5	634	28.0	33.7	5.7	2564	58.1	56.9	-1.2
	6	513	24.9	32.1	7.2	2692	56.2	56.4	.2
	7	453	17.7	26.4	8.7	2625	57.8	59.0	1.2
	8	456	6.2	16.9	10.7	2759	57.2	59.0	1.8
MATHEMATICS	4	724	18.6	30.6	12.0	2659	50.7	53.3	2.6
	5	636	18.6	29.6	11.0	2561	54.0	54.3	.3
	6	579	13.0	26.7	13.7	2701	53.1	53.4	.3
	7	456	6.6	19.2	12.6	2645	54.2	54.5	.3
	8	458	9.2	19.2	10.0	2795	54.8	55.1	.3

[a]No data were available for grade three.

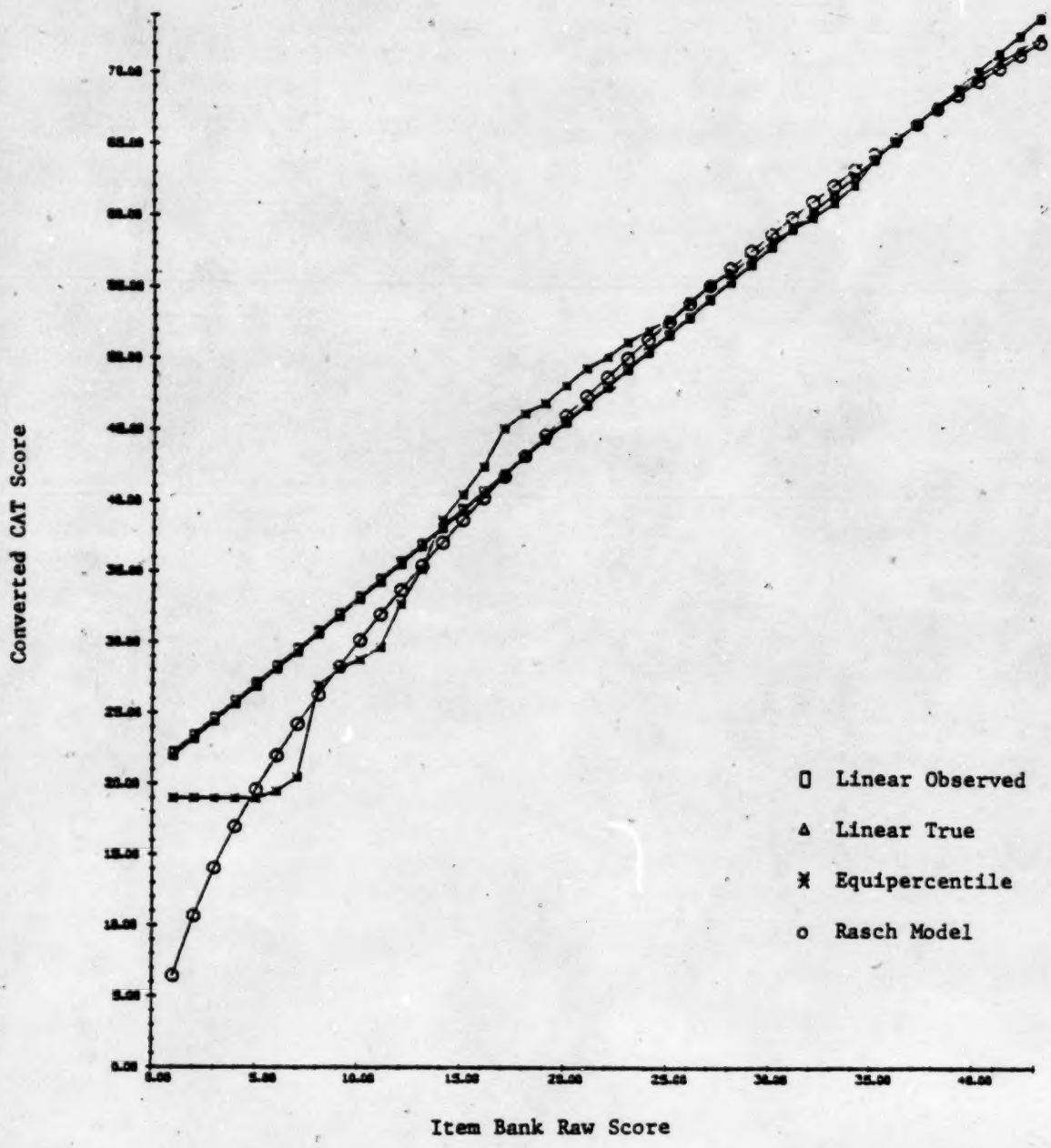

Figure 1

Grade Three Converted CAT Reading Scores Yielded by Four Equating Methods for Scores on the Corresponding Item Bank Test.

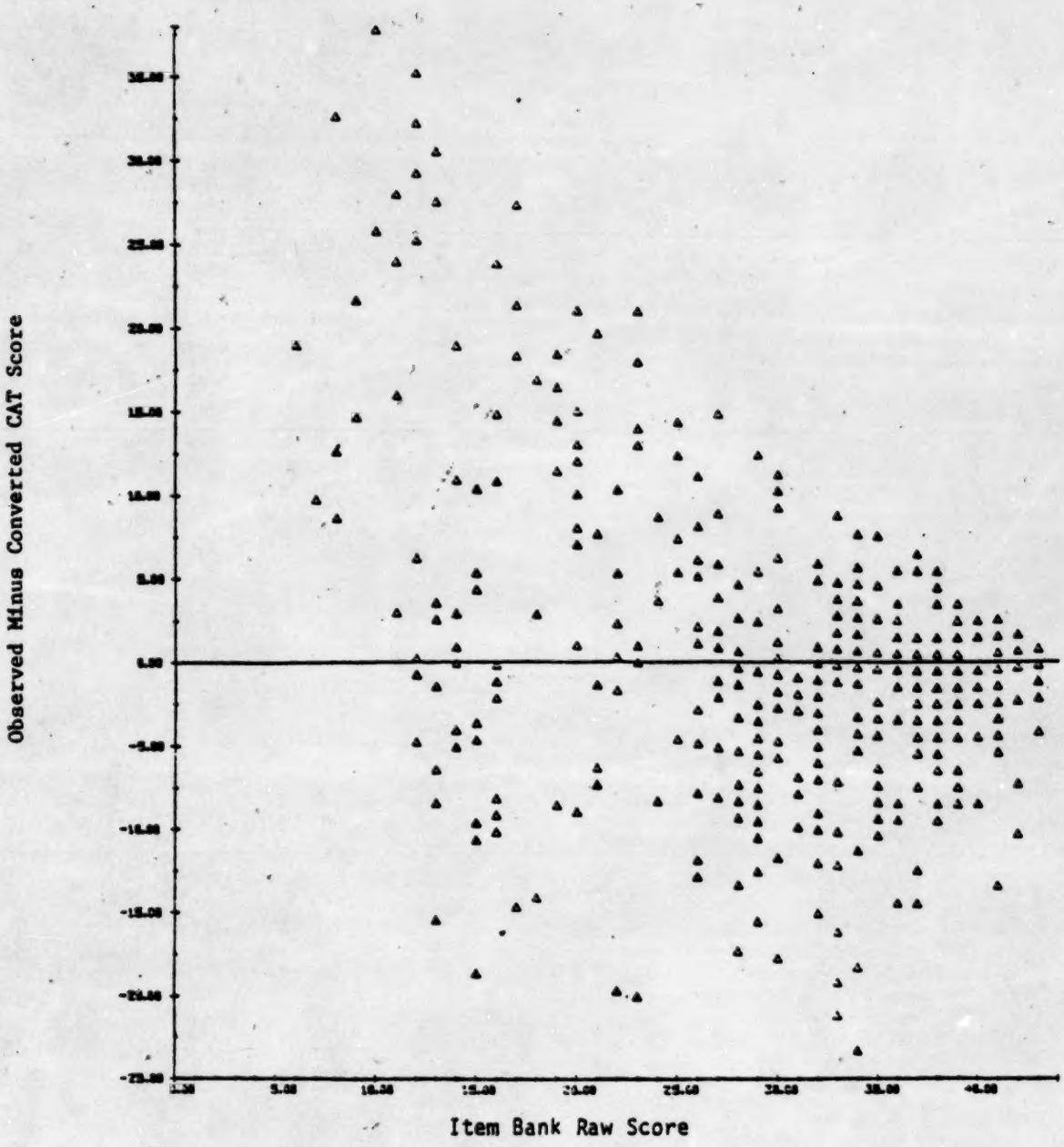

Figure 2

Grade Three Discrepancy Between the CAT Reading Score Actually Observed for Each Student Tested and the Converted CAT Score Yielded by the Rasch Model Equating Technique by Item Bank Score.